Eliciting and aggregating individual expectations: An experimental study

Maastricht University

Ronald Peeters, Leonard Wok

Eliciting and aggregating individual expectations: An experimental study

RM/14/029

GSBE

Maastricht University School of Business and Economics
Graduate School of Business and Economics

P.O Box 616
NL- 6200 MD Maastricht
The Netherlands

Eliciting and aggregating individual expectations: An experimental study*

Ronald Peeters† Leonard Wolk‡

July 21, 2014

Abstract

In this paper we present a mechanism to elicit and aggregate dispersed information. Our mechanism relies on the aggregation of intervals elicited using an interval scoring rule. We test our mechanism by eliciting beliefs about the termination times of a stochastic process in an experimental setting. We conduct two treatments, one with high and one with low volatility. Increasing the underlying volatility affects the location of the interval, yet it does not significantly affect its length. Consequently, individuals perform significantly better in the low volatility treatment than in the high volatility treatment. Next, we construct distributions by aggregating intervals across different individuals. Our results reveal that the predictive quality of the aggregated intervals (as measured by the Hellinger distance to the true distribution) increases by more than 30% when increasing the aggregation level from two to eight individuals. This shows that aggregating individual intervals may be an attractive solution when market mechanisms are infeasible.

JEL Classification: C53, D84, C91.

Keywords: forecasting, belief elicitation, information aggregation, experimental economics.

*We thank the audiences at the Borsa Istanbul 2014 Workshop on Behavioral Finance, the INFORMS Annual Meeting 2013 in Minneapolis, the Maine Economics Conference 2014 in Waterville, the University of Paderborn, as well as Matt Embrey, Ben Gillen, Jörg Gross, Stephan Smeekes, Sasha Vostroknutov and Maria Zumbühl; Ronald Peeters gratefully acknowledges funding from NWO.

†Department of Economics, Maastricht University. E-mail: r.peeters@maastrichtuniversity.nl.

‡Department of Finance, Maastricht University. E-mail: l.wolk@maastrichtuniversity.nl.

1 Introduction

Firms often depend on internally generated forecasts when making operational decisions such as whether to invest in a project or whether to increase production capacity. Generating forecasts for such purposes requires both collection and aggregation of information that is dispersed across different individuals within, as well as outside, the firm. Given that unstructured mechanisms to aggregate information may result in a failure to correctly take all information into account (Hopman, 2007) and given the importance of having access to accurate information when making decisions, information aggregation of subjective data has received surprisingly little attention in the literature. In this paper we propose and implement a non-market based mechanism to aggregate information and to test it experimentally. There are several advantages to moving away from a market-based setting. For instance, our approach can be operated with fewer forecasters, individual forecasts can be conditioned on the individual characteristics of the specific forecaster and information flows can easily be traced across different subsets of the forecasters.

In our experiment, each subject has to forecast over a sequence of twenty periods when a time series will hit either a predefined upper or lower bound. The parameters of the random process from which the time series are generated are fixed and subjects gradually learn about the underlying parameters when the experiment advances. Our experiment does not involve strategic interaction, i.e. there is no competition among subjects and they are rewarded purely based on their own performance.

We believe that several aspects of our design are novel. First, we do not only elicit point predictions but also intervals using an interval scoring rule as proposed by Schlag and van der Weele (2009). Using an interval scoring rule, in contrast to confidence intervals, allows an individual to reveal her forecast without the need to reason about probabilities. Relatedly, it has been shown that eliciting quantiles, instead of probabilities, generate better forecasts with lower variance (Lichtendahl et al., 2013). Second, we aggregate the elicited intervals to construct forecasted distributions and compare them to the 'true' distribution of the data generating process. Our method allows us to compare the aggregated distributions to the 'true' distribution and hence allows us to judge the quality of the forecasted distribution, and we use the Hellinger (1909) distance as a measure to do so.

Our findings show that on the individual level the interval predictions do not change significantly over time. Individuals do not appear to maximize expected payoffs when constructing intervals. Instead, subjects seem to maximize expected payoffs conditional upon choosing an interval length, i.e. they fine-tune the location of the interval given their preference for a certain length of this interval.

Furthermore, we find that when aggregating individual intervals into distributions, the marginal improvement is positive, statistically and economically significant. The marginal effect is positive but declining with the number of intervals used in the aggregation process. For instance, moving from one to two individuals, improves the performance of the forecast by approximately 55%, while moving from eleven to twelve individuals yields an improvement of about 5%. Interestingly, while on an individual level participants in the high volatility treatment perform worse than in the low volatility treatment, when aggregating intervals this finding reverses: the aggregated forecast in the high volatility treatment is significantly better than the one in the low volatility treatment. It appears that the high volatility treatment creates more variance in the individual intervals that better fits the tails of the distribution when aggregating them and thereby reduces the Hellinger distance.

The findings of our study show that non-market based mechanisms can be used successfully for forecasting. Specifically, the elicitation of intervals using scoring rules are informative to a decision maker, especially after aggregating several intervals into distributions. Our findings corroborate the findings by Gillen et al. (2013), who study sales forecasting within Intel using a non-market based mechanism. These authors show that their mechanism performs well, and when compared to internal sales forecasts, the mechanism outperforms in a majority of the cases. Goel et al. (2010) also study non-market based methods and show that they do not perform significantly worse than a prediction market. We offer complementary evidence from an experimental setting, which has as advantage that we know the underlying process from which the outcomes are generated, and hence we can compare the forecasts with the true distribution rather than only with the realization itself, i.e. the experimental setting facilitates a better performance analysis.

In the remainder of this paper we outline the design and provide the results of our experiment. The paper is organized as follows. In the second section we review the literature relevant to our experiment, and in the third section we describe the experimental set-up in detail. In section four we present the results of the study and finally, in section five, we discuss the findings of the study and conclude.

2 Related literature

Forecasting future events has received attention from several streams of literature in different disciplines and can broadly be divided into two different groups. On the one hand, a large literature has emerged on prediction markets as well as experimental double auctions. These studies focus on the market mechanism, in the form of a double auction, and its ability to aggregate dispersed information. The other strand of literature focuses on the elicitation of

beliefs and centers around the use of scoring rules. In this section we review these strands of the literature and connect them to contributions of our paper.

Prediction markets have received much attention in the literature and have shown to forecast many different events such as presidential elections very accurately (Forsythe et al., 1992). Several attempts have been made to implement and use these markets in corporate settings (Chen and Plott, 2002; Cowgill and Zitzewitz, 2013), often with very positive results. These results are in line with the insights derived from experimental asset markets. Smith (1962), Plott and Sunder (1982) and Forsythe et al. (1982), for example, show that information aggregation works well in the laboratory and that prices converge to a rational expectations equilibrium. Further evidence by Plott and Sunder (1988) suggests that when market institutions are well designed, these markets perform according to rational expectations. While this is strong evidence in favor of double auction based markets in general, Bossaerts and Plott (2004) extend the analysis to a risk and return framework, the capital asset pricing model. These authors show that, when the market is thick enough, assets are priced correctly. This is however not the case when markets are thin in which case they do not adjust fully to reflect the fundamental asset price (Bossaerts and Plott, 2002). In the case of a prediction market, this result shows that it is difficult to make inference on the probability of an event when there is insufficient trading. This finding, is further elaborated upon by Healy at al. (2010), who show that when markets are thin, iterated polls may outperform the double auction mechanism.

Despite the successes of prediction markets, they have not received much support from legal institutions, which have made it very difficult to operate them with monetary incentives (Arrow et al., 2008), an often essential feature to ensure that the right incentives are present. A further problem has been the fear of manipulation which potentially can go undetected and lead to wrong forecasts. The evidence is mixed, and while Hanson et al. (2006) find that manipulators often are unable to distort prices, Veiga and Vorsatz (2009) find that manipulators are able to affect prices and sometimes even earn positive profits. Overall, this shows that especially when markets are thin, and not enough forecasters exist as traders, it may be difficult to use a double auction mechanism as a forecasting tool. A further shortcoming is that while a market mechanism aggregates information into a price, it does not reveal the distribution around the forecast. Our approach reveals not only the predicted mean but also the distribution around it. This gives the decision maker a better understanding of the higher moments around the expected value, which can be used to construct better forecasts.

A well developed alternative is to elicit beliefs by the use of scoring rules. Scoring rules assign a score to a forecaster that depends on the predicted outcome as well as the actual outcome. The scores can be used to evaluate forecasts as well as forecasters. A forecaster

who is consistently better than another one, in terms of his scores, can be considered better (Winkler, 1971). Relying on the actual scores to evaluate the quality of a forecaster may not always be feasible, for instance for events that only occur once such as the start of a war, or in situations where the incentives provided by the scoring are not enough to keep the forecaster honest. Under such circumstances, the Bayesian Truth Serum (BTS) provides a good adaptation of the logarithmic scoring rule (Prelec, 2004). BTS does not depend on a common prior by all forecasters. Instead, forecasts are scored by their information score of the actual-to-predicted ratio as well as their prediction score. The prediction score is based on how well the forecaster is able to predict the answers of the average forecaster. Thus, the score for each forecaster depends on his own forecast as well as the ability to predict the forecasts of other forecasters. The aggregate score for each forecaster thus provides the decision maker with important information about their ability to forecast the beliefs of others and can help the forecaster determine the weight he places on each of the individual forecasts. The drawback of this method is that it relies on a rather large number of forecasters. In this paper, we are concerned with environments where the number of forecasters are limited and it is not feasible to implement the Bayesian truth serum.

Recently, the problem of the number of forecasters required has received considerable attention in the area of the 'wisdom of crowds'. The idea is that if we can identify experts that outperform the average forecaster, we can generate better and more accurate forecasts with fewer forecasters. Budescu and Chen (2014) approach this problem by defining the share of the predictive performance that can be attributed to each forecaster and then weighing forecasts based on this measure. Goldstein et al. (2014) confirm these findings and show that it is possible to identify a subset of players in advance in a fantasy soccer league that outperform the average. Thus, performance appears to be persistent over time.

In this paper, we develop an alternative mechanism to the ones proposed above. In the next section we describe our elicitation method, that is also based on scoring rules, in detail.

3 Experiment

3.1 Setting

In this experiment, participants are exposed to a random process. The random process starts at a value of zero at time $t = 0$ and runs from there in discrete time-steps. Each unit of time the value is incremented with a randomly drawn number (possibly negative) that is drawn according to a normal distribution with mean zero (hence, there is no drift). The random process terminates either when the value has dropped below the lower boundary at -2.5 or has increased to above the upper boundary at $+2.5$, or has reached time $t = 100$ without

having touched or crossed one of the boundaries. Figure 1 contains one example of such a time series that terminated at time $t = 63$.

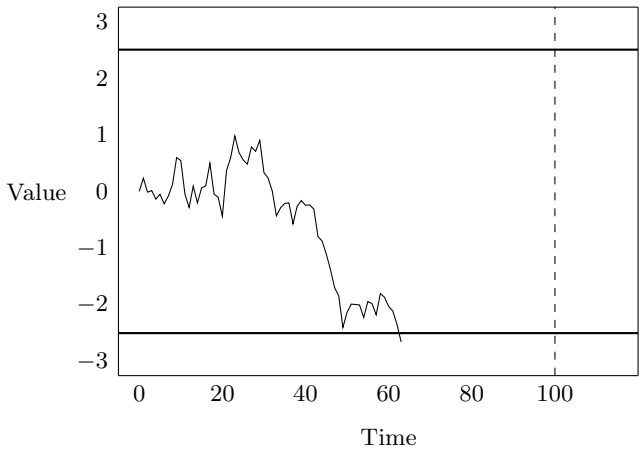


Figure 1: An example of a time series.

We implement two different processes: one with a low standard deviation of 0.1885 (the low volatility treatment) and one with a high standard deviation of 0.2270 (the high volatility treatment). These standard deviations are chosen such that the probability of the process to terminate before $t = 100$ equals 1/3 in the low volatility treatment and 2/3 in the high volatility treatment. Figure 2 presents the (cumulative) distribution over termination times conditional on termination before $t = 100$ for the two treatments. It can be seen in the right panel that the distribution of the low volatility treatment first-order stochastically dominated that of the high volatility treatment.

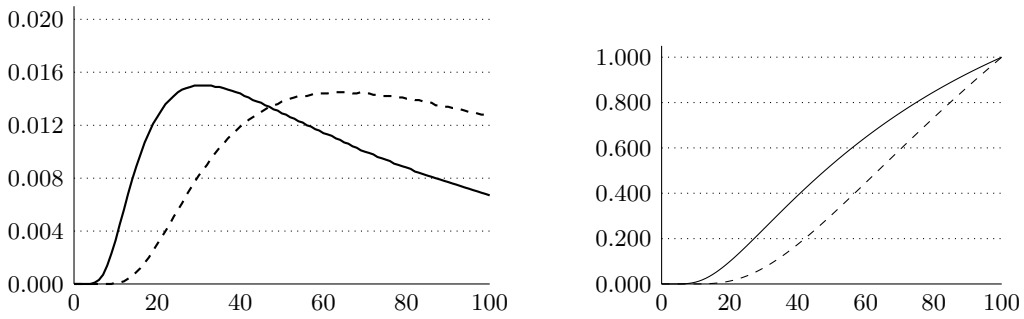


Figure 2: Distribution of termination times conditional on termination before $t = 100$. The dashed curves relate to the low volatility treatment; the solid curve to the high volatility treatment. The left panel shows the density; the right panel the cumulative distribution.

3.2 Two prediction tasks

Participants were faced with two prediction tasks regarding the termination time of the time series to be generated by the random process. First, they were asked the following question:

> How likely do you regard the event that the time series is going to hit the boundary before time $t = 100$?

Participants had to indicate their answer by positioning a triangular cursor on a line of which the extreme points corresponded to the answers "totally unlikely" ($z = 0$) and "totally likely" ($z = 100$), as is shown in Figure 3. Participants were incentivized in accordance to the

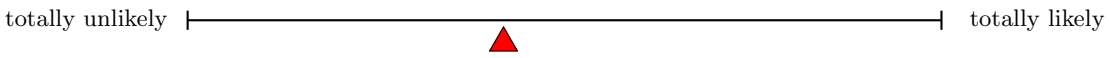


Figure 3: Point prediction.

quadratic scoring rule, an incentive compatible mechanism to elicit beliefs over discrete events (cf. Brier, 1950; Offerman et al., 2009). A participant expressing the expectation $\hat{z} \in [0, 100]$ received

$$50 + 2 \cdot 50 \cdot \tfrac{\hat{z}}{100} - 50 \cdot [(\tfrac{\hat{z}}{100})^2 + (1 - \tfrac{\hat{z}}{100})^2] \quad \text{ECU}$$

if the contested event was realized and

$$50 + 2 \cdot 50 \cdot (1 - \tfrac{\hat{z}}{100}) - 50 \cdot [(\tfrac{\hat{z}}{100})^2 + (1 - \tfrac{\hat{z}}{100})^2] \quad \text{ECU}$$

otherwise. While moving the triangular cursor along the line, the potential payoffs in either event were shown on-screen in real-time.

Second, we gave them the following task:

> Conditional on the time series hitting the boundary before time $t = 100$, indicate the time interval in which you believe it is going to hit the boundary.

Participants expressed their beliefs by positioning two triangular cursors on the time line between $t = 0$ and $t = 100$: one of the cursors indicated the lower bound (x) of the interval, the other the upper bound (y); see Figure 4. Participants were incentivized by means of an

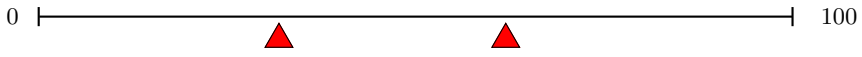


Figure 4: Interval prediction.

interval scoring rule, a mechanism that provides information about both the location and the dispersion of the true belief distribution in case of single-peaked beliefs (see Schlag and

van der Weele, 2009). A participant expressing the belief that, conditional on the time series hitting the boundary before $t = 100$, it to hit the boundary in the interval $[\hat{x}, \hat{y}]$ received

$$100 \cdot (1 - \tfrac{\hat{y}-\hat{x}}{100})^2 \quad \text{ECU}$$

if the time series indeed terminated in the given interval and 0 ECU otherwise. Note that the potential payoff that could be obtained is larger when a smaller interval is selected. Again, the potential payoff was shown on-screen in real-time when the cursors were moved along the time line.

3.3 Procedures

A random selection of subjects from our subject pool (mainly students in business and economics) were invited and could sign up to participate in one of two sessions of an economic experiment via ORSEE (Greiner, 2004). The sessions were run in the BEElab at Maastricht University in September 2013. The instructions were paper-based and the prediction phase was computerized using z-Tree (Fischbacher, 2007).

A session consisted of twenty rounds. Before the first round started, participants were shown an animation of a time series that was generated from one of the given random processes. After having seen this animation, they were confronted with the two prediction tasks to forecast the outcome of the next time series that was generated from the same random process.[1] In order to avoid the unfortunate event of accidental decision making, participants had to confirm their decisions by ticking a little box. After having confirmed their predictions, participants were shown the time series that was generated for the first round. After having seen that animation, they had to give their expectations regarding the outcome of the time series for the second round. This procedure continued until the last (twentieth) round. As the random process from which all time series were generated were kept fixed during the entire session, over rounds participants gradually became more familiar with the underlying process. In order to make aggregation of individual predictions sensible, all participants in a treatment were shown the same animations in the same order.

At the end of the session, for each participant individually, eight random draws (with replacement) over the payoffs that were earned over the two tasks in the twenty rounds were made. The final earnings of the participants consisted of the amount of ECUs collected in these eight tasks exchanged into Euros (at a conversion rate of 6 Eurocents for each ECU) and a 3 Euro show-up fee. Finally, the participants participated in a short cognition task in which we measured their perceptual reasoning ability and we elicited their risk attitude. For

[1] The time series were generated by a statistical software package and were not manipulated for the purpose of this experiment.

the cognition task, we used the symbol-digit correspondence test from the Wechsler Adult Intelligence Scale (WAIS), in which subjects had 90 seconds to find as many correspondences between symbols and numbers as they could, using the correct number for each symbol. Speed and accuracy under time pressure determine an individuals ability (cf. Dohmen et al., 2010). Risk attitude was elicited by the direct approach as suggested in Dohmen et al. (2011). Moreover, we elicited a few personal characteristics, including gender and age.

4 Results

In total 48 students participated in the two treatments with an even division between the high and low treatments. Each experimental session lasted about 60 minutes and the average earnings of the subjects was 13.56 Euro.

The top part of Table 1 shows the summary statistics of the main characteristics of the participants in our experimental sessions. The ratio of males was slightly larger in the low volatility treatment; so was the number of correctly identified symbols in the cognition task. There are no substantial difference in age and risk attitude (where the value 0 indicates extreme risk aversion and the value 10 extreme risk loving) between the participants in the two treatments. The bottom part of this table shows average decisions taken over all individuals over all twenty periods. Overall, the participants in the low volatility treatment are significantly better at the interval forecasting task than the participants in the high volatility treatment.

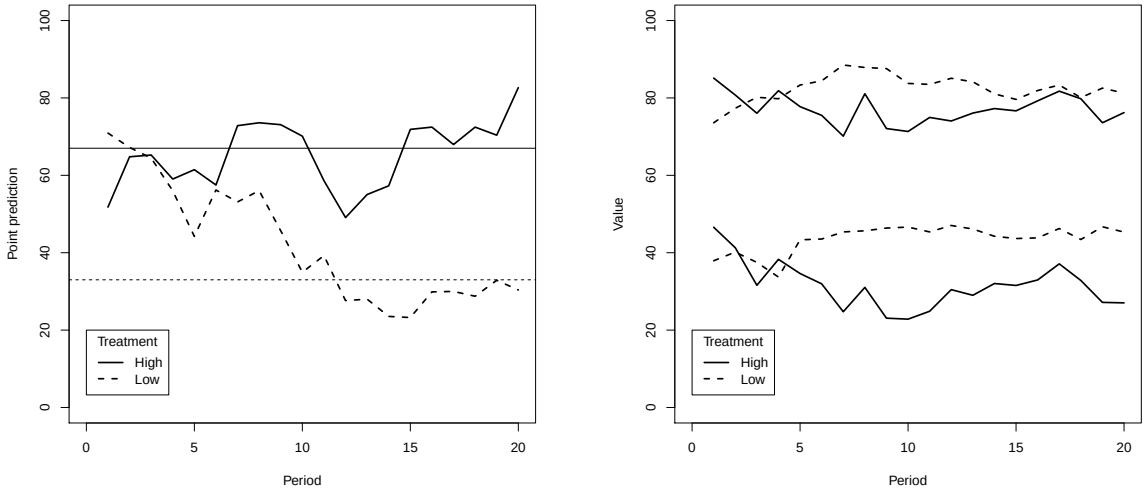
	Mean value		
	All	Low	High
Age (years)	21.2	21.2	21.3
Gender (%, Male = 1)	50.0%	58.3%	41.7%
Risk attitude (0–10)	6.1	6.0	6.1
Cognitive ability (number)	40.5	41.1	40.0
Point prediction (0–100)		42.1	65.4
Lower bound (0–100)		43.6	31.6
Upper bound (0–100)		82.4	77.1
Point prediction (exp. payment)	35.2	34.6	35.9
Interval prediction (exp. payment)	16.4	18.2	14.6

Table 1: Summary statistics of the participants in the experiment.

The following sections discuss the results of the experiment. We start with the point and interval predictions at the individual level and we subsequently discuss the quality of the forecasts when we aggregate the intervals of different participants.

4.1 Individual predictions

Figure 5 presents average subject decisions over time. The left panel shows that initially, after seeing two sample series (one in the instructions and one on screen), subjects regard the likelihood that the next time series will terminate before $t = 100$ not too different in the two treatments and that over time the difference in predictions becomes visible. In the low volatility treatment the average prediction drops to the true probability of $1/3$ in about twelve decision periods in order to fall a bit below this probability from there onwards. In the high volatility treatment the average prediction oscillates around the true probability of $2/3$.



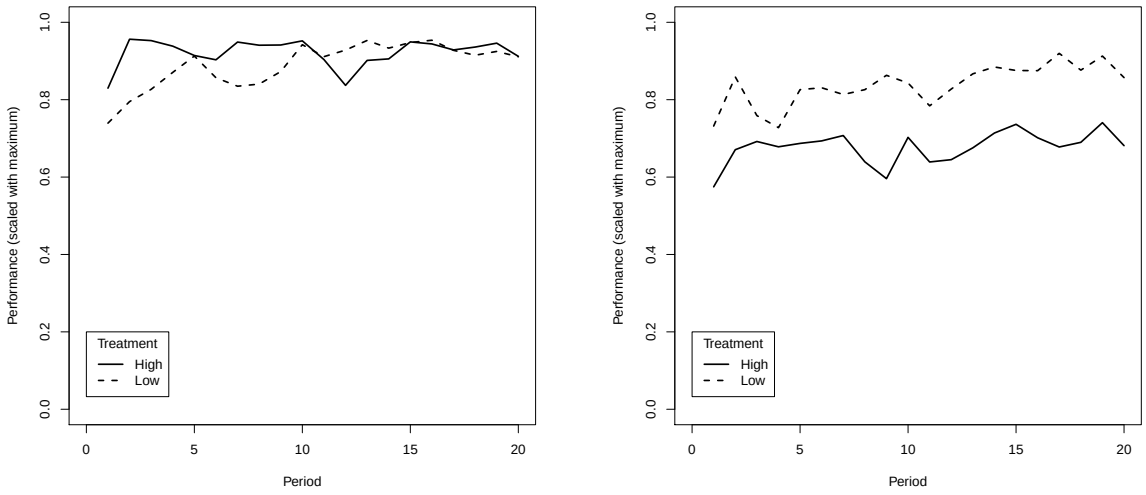
(a) Point prediction.

(b) Interval prediction.

Figure 5: Average decisions by the subjects over time.

The right panel shows the average lower and upper bounds in the two treatments over time. Given the true distributions on termination times conditional on termination before $t = 100$ (see Figure 2), the interval that maximizes the expected payoff is $[51, 83]$ in the low volatility treatment and $[21, 51]$ in the high volatility treatment. So, the payoff maximizing intervals do not show any overlap and are rather similar in length. The figure shows that in the low volatility treatment the average prediction is close to the expected payoff maximizing interval from the fifth period onwards. Also in the high volatility treatment learning effects are strongest in the first five periods, but the average prediction underestimates the termination time and indicates the use of a larger interval compared to the payoff maximizing one. By design of the experiment we would expect that the intervals would not overlap to such a large extent, and it appears as if individuals are largely invariant to the underlying volatility in this experiment.

Next, we investigate how performance develops over time in our experiment. In order to properly assess the quality of the individual decisions given the incentives provided, Figure 6 presents the average expected payoffs (given the true probabilities and probability distributions), relative to the maximum possible payoff. The figure thus shows how far away the subjects are from the optimal decision if she had known the underlying data generating process of the time series. Panel (a) shows the average performance for the point predictions and panel (b) shows the average performance for the interval predictions.



(a) Point prediction

(b) Interval prediction

Figure 6: Average performance of the subjects over time.

For the point predictions the average subjects' performance improves over time in the low volatility treatments and reaches a level close to the maximum; for the high volatility treatment there is not such a clear trend, partly because of the high performance attained already in the second period. For the interval predictions the average performance improves in both treatments, but in absolute terms the performance in the high volatility treatment lags systematically behind that in the low volatility treatment and does not catch up over time. Thus, on an individual level, estimating termination times of a random process appears to be more difficult when the process is more volatile.

Table 2 presents the results of the regressions where we regress individual forecasting performance against treatment and individual characteristics. For the sake of exposition we divide the twenty time periods into five equally sized terms (1–4, 5–8, 9–12, 13–16 and 17–20) and we control control for (treatment-specific) learning via these term dummies. The dependent variable in the three columns are, respectively, performance in the point prediction task, performance in the interval prediction task, and performance in the interval prediction task

conditional on the chosen length of the interval. So, the third column measures performance relative to the location of the interval given the length of the interval that a participant chose. This rescaling in performance enables us to segregate the effects of the choice of length and of location on performance.[2]

			Interval			
	Point		(1)		(2)	
Constant	0.89593^{***}	0.03625	0.84664^{***}	0.04890	0.92172^{***}	0.02622
Term 2	0.05323^{**}	0.02534	0.05470^{**}	0.02768	0.03017^{*}	0.01605
Term 3	0.10560^{***}	0.02213	0.06000^{**}	0.02945	0.03982^{***}	0.01419
Term 4	0.13896^{***}	0.01927	0.10603^{***}	0.02480	0.04250^{***}	0.01497
Term 5	0.11158^{***}	0.02070	0.12215^{***}	0.02518	0.04650^{***}	0.01547
Treatment	0.10998^{***}	0.02250	-0.11025^{***}	0.02948	-0.11681^{***}	0.02161
Term 2 × Treatment	−0.04577	0.02994	−0.02692	0.03758	0.05332^{**}	0.02593
Term 3 × Treatment	-0.11635^{***}	0.02912	-0.06831^{*}	0.03903	0.03698	0.02502
Term 4 × Treatment	-0.13313^{***}	0.02512	−0.05328	0.03470	0.04565^{*}	0.02366
Term 5 × Treatment	-0.10016^{***}	0.02551	-0.07890^{**}	0.03520	0.04580^{*}	0.02421
Gender	-0.01819^{**}	0.00832	0.01271	0.01153	0.00130	0.00764
Risk attitude	-0.01314^{***}	0.00220	-0.01857^{***}	0.00298	−0.00058	0.00176
Cognitive ability	0.00004	0.00068	0.00065	0.00087	0.00024	0.00047
Adjusted R^2	0.1262		0.2375		0.1696	
Number of obs.	960		960		960	

Table 2: Cross-sectional regressions of participants' characteristics on the expected payoff in the experiment. ***.001, **.01, *.05.

The results for the point prediction task shows that throughout both treatments subjects do better in all terms following the first one. First term performance is significantly higher in the high treatment, but the effect disappears in the third term. Moreover, we see that females and more risk averse subjects perform better in this task, but that cognitive ability has no significantly affect on performance.

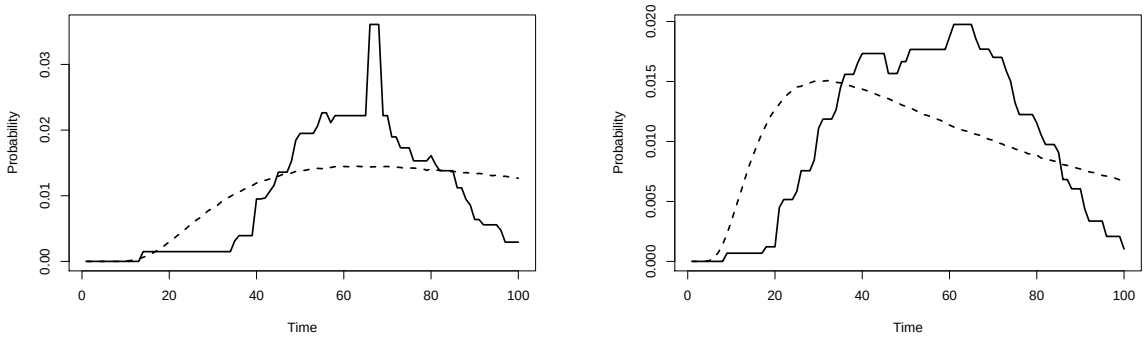
Further, we see that performance in the interval prediction task improves over time and subjects are significantly better in all terms following the first one. As we saw already in Figure 6, performance is significantly worse in the high treatment. Also in this task cognitive ability has no significant impact and is risk aversion found to destroy performance; gender, though, has no impact in this task. When we measure performance conditional on the length of the chosen interval (third column), we see that the negative effect of high volatility on performance remains similar in magnitude and significance. However, the interaction of treatment with the term dummies are now significantly positive, indicating that subjects in the high treatment are improving the location of the interval over time. Once we control for the length of the interval, the significance of the role of risk attitude disappears, which hints

[2]Note that although the underlying random process is driftless, participants may perceive a drift from the realized time series. The reported results do not change when we control for the 'perceived' drift that they may have observed.

at the earlier significant effect of risk attitude being mainly working via the length and not the position of the interval.

4.2 Aggregation of individual predictions

The aggregation of interval predictions of several subjects yields a distribution over possible termination times. Such an amalgamation of individual forecasts may provide a better forecast than any of the individual forecasts. Figure 7 shows the aggregated probability density functions for the two treatments in the last period of the experiment, where for each treatment the aggregation is taken over all 24 participating subjects. We focus on the quality of an aggregated prediction in relation to group size, and how the quality develops over time.



(a) Low volatility treatment

(b) High volatility treatment

Figure 7: Probability density functions of the predicted distribution (solid line) and the true (dashed line).

In order to study the impact of group size (and composition) on the quality of predictions when aggregating individual predictions over groups it is important to adopt a good measure to quantify 'quality of prediction'. One property that such a measure should capture is that it allows for a fair comparison within and across groups of different sizes. In our analysis, we will make use of the *Hellinger distance* (Hellinger, 1909) that quantifies the similarity between two probability distributions. An important advantage of the Hellinger distance over often used alternatives (such as the Kullbeck-Leibler divergence) is that it does not require absolute continuity, a property that is violated almost by design.

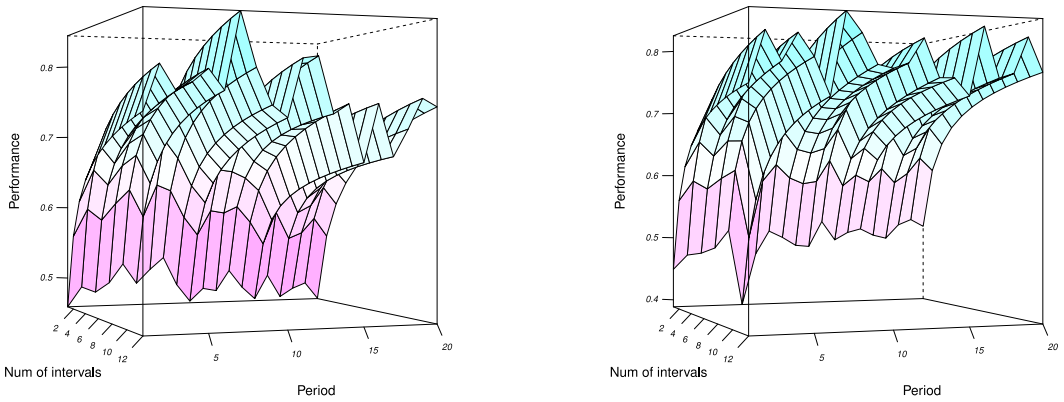
The Hellinger distance of the (discrete) empirical probability distribution $Q = (q_1, \ldots, q_m)$ to the (discrete) true probability distribution $P = (p_1, \ldots, p_m)$ is defined as

$$H(Q, P) = \frac{1}{\sqrt{2}} \sqrt{\sum_{j=1}^{m} (\sqrt{q_j} - \sqrt{p_j})^2}.$$

In the case the two distributions P and Q coincide, the Hellinger distance equals zero. The maximum Hellinger distance of one is obtained when the supports of the two distributions are disjoint. Consequently, for intuitive reasons, we henceforth define a performance index, Z, that equals one minus the Hellinger distance:

$$Z(Q, P) = 1 - H(Q, P).$$

In Figure 8 we plot the performance measure, Z, of the aggregated interval predictions over different group sizes and time periods. In the three dimensional graph, each point represents the average performance for a given aggregation size (increasing from far to near) and time period (increasing from left to right). The left panel shows this for the low volatility treatment and the right panel for the high volatility treatment. The graphs from both treatments look quite similar and it is evident that the performance improves substantially when increasing the group size. In both treatments, for given group size, the performance averaged over all possible groups of that size is rather constant over time. Any effects of learning that we saw to be present on the individual level, in particular during the first eight periods, seem to have disappeared in the aggregation process.



(a) Low volatility treatment

(b) High volatility treatment

Figure 8: Average performance of interval predictions over group size and over time.

While we observe a clear improvement in the performance when increasing the number of individuals in a group in the two graphs, next we estimate the following linear regression model where we can control for the individual heterogeneity across different group combinations:

$$\begin{aligned} Z_{gt} &= \beta_0 + \beta_1\, Treatment + \beta_2\, Gender_g + \beta_3\, RiskAttitude_g + \beta_4\, CognitiveAbility_g \\ &\quad + \textstyle\sum_{t=2}^{T} \gamma_t\, D_t + \sum_{s=2}^{S} \kappa_s\, D_s + \sum_{i=2}^{I} \lambda_i\, D_i + \varepsilon_{gt}. \end{aligned}$$

Here, Z_{gt} is the average performance of group configuration g in term t. The reason to restrict to terms is to reduce the computational burden of solving the least-squares problem. D_t and D_s capture respectively the term and the group size effect; so D_s takes a value of one if group g is of size s. We further control for treatment effect, gender composition in the group, average risk attitude in the group and average cognitive ability in the group. In addition we control for individual performance effects using individual dummy variables; D_i takes a value of one in case individual i is a member of group g.

With the setup of five terms, twenty-four individuals in each treatment in combinations (i.e. group sizes) of at most twelve subjects as well as two treatments, this results in 97,406,850 observations. To solve for the coefficients of this model we wrote a program in C++ using the GNU Scientific Library (GSL).[3] The results are reported in Table 3.

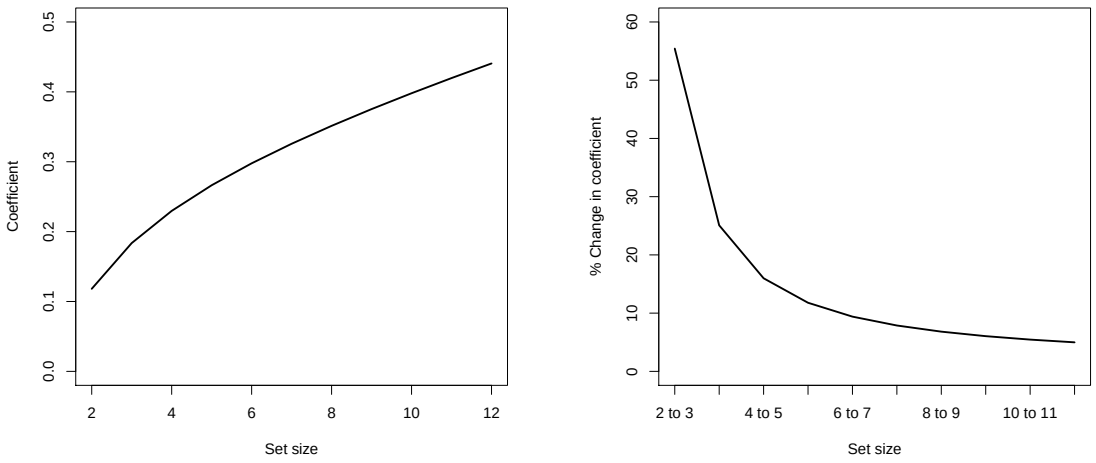
	Group performance	
Constant	0.51808^{***}	0.00147
Group size 2	0.11812^{***}	0.00148
Group size 3	0.18359^{***}	0.00143
Group size 4	0.22963^{***}	0.00142
Group size 5	0.26634^{***}	0.00142
Group size 6	0.29772^{***}	0.00142
Group size 7	0.32570^{***}	0.00142
Group size 8	0.35135^{***}	0.00142
Group size 9	0.37531^{***}	0.00142
Group size 10	0.39800^{***}	0.00142
Group size 11	0.41971^{***}	0.00142
Group size 12	0.44064^{***}	0.00142
Term 2	0.01965^{***}	0.00001
Term 3	0.00102^{***}	0.00001
Term 4	-0.02618^{***}	0.00001
Term 5	-0.02357^{***}	0.00001
Treatment	0.00788^{***}	0.00004
Group gender (mean)	-0.00142^{***}	0.00011
Group risk attitude (mean)	0.00901^{***}	0.00003
Group cognitive ability (mean)	-0.00166^{***}	0.00002
Individual effects	Yes	
R^2	0.6570	
Number of obs.	97,406,850	

Table 3: This table shows the coefficients and the associates standard errors for the aggregated performance regression. The significance levels indicated in the table are defined as follows: $^{***}.001$, $^{**}.01$, $^{*}.05$.

When we increase group size, the predictive performance of the group improves, yet at a diminishing rate. In Figure 9(a) we plot the group size coefficients and in Figure 9(b) we plot the marginal improvement of the current coefficient over the previous group size coefficient. The improvement is statistically and economically significant at each level of group size,

[3]Available as an open-source library at `https://www.gnu.org/software/gsl/`.

yet it levels off after about eight individuals in a group. This result shows that increasing the number of forecasters improves the performance of the prediction and can help decision makers make better predictions. Interestingly, there is very little difference in the forecasting performance between the two treatments and the coefficient is approximately two orders of magnitude lower than the effect of an increase in the group size. Thus, the underlying volatility of the time series in our experiment does not directly appear to influence performance of aggregate predictions. While our control variables, gender, risk attitude and cognitive ability are statistically significant (due to the large sample size), the economic significance is very small. Of these, the most striking effect may be the one for risk attitude. Where on the individual level more risk averse individuals perform better, when aggregating on group level, groups consisting of more risk averse individuals, on average, perform worse. The reason for this may be that risk averse individuals take larger intervals,[4] and that the aggregation of multiple intervals leads to a better distribution when these intervals are smaller.



(a) Performance and set size

(b) Change in performance

Figure 9: This figure shows the coefficients of the group size dummies from the linear regression presented in Table 3. Panel (a) shows the value and Panel (b) shows the percentage change when increasing the group size.

5 Discussion and conclusion

We introduce a novel mechanism to generate forecasts in a laboratory setting with the advantage of observing the data generating process and thereby we can compare forecasts to the underlying distribution instead of to an actual outcome that is drawn from an unknown

[4]Recall here that the significance of the effect of risk attitude on individual performance in the interval prediction task disappeared when controlling for interval length.

distribution.

Our results show that information aggregation can help improve forecasts also when the aggregation is done outside a market-based setting. In particular, there are large gains to obtain from aggregating individual interval forecasts into distributions. This means that even with a small number of potential forecasters, a decision maker can generate informative forecasts in order to aid decision making in many different contexts. In particular, we show that in our experimental setting possible gains arising from learning over time are small in comparison to the effects of aggregation over multiple forecasters. This effect is economically large and demonstrates the power of information aggregation.

Several implications arise from our study. First, we show that using scoring rules as a mechanism to elicit beliefs is viable and provides a flexible alternative to market-based mechanisms. Second, we show that in particular interval forecasts elicited using the scoring rule proposed by Schlag and van der Weele (2009) works well. In particular, subjects appear to do better on an individual level when the variance is low. Yet, when we aggregate the individual intervals into distributions the aggregate distributions generated in the high variance treatment generates a better quality forecast (as measured by the Helling distance). This is due to the larger variance of the individual intervals in this high variance treatment, and thus when we aggregate the intervals we get a better fit in the tails. Third, individual attributes such as gender, risk attitude and cognitive ability appear to play a small role in the performance on both the individual and the aggregate level.

Overall, the results show that by better understanding both elicitation and aggregation of beliefs into forecasts we can improve forecasting in settings where information is dispersed and not easily communicated in an unbiased way. By developing and testing such mechanisms in the laboratory we can improve our understanding of how to create even better mechanisms and generate mechanisms to be employed in the field.

References

1. Arrow, K. J., R. Forsythe, M. Gorham, R. Hahn, R. Hanson, J. O. Ledyard, et al. (2008). The promise of prediction markets. Science 320(5878): 877-878.

2. Bossaerts, P. and C. Plott (2002). The CAPM in thin experimental financial markets. Journal of Economic Dynamics and Control 26: 1093-1112.

3. Bossaerts, P. and C. Plott (2004). Basic principles of asset pricing theory: Evidence from large-scale experimental financial markets. Review of Finance 8: 135-169.

4. Brier, G. (1950). Verification of forecasts expressed in terms of probability. Monthly

Weather Review 78: 1-3.

5. Budescu, D. V. and E. Chen (2014). Identifying expertise to extract the wisdom of the crowds. Management Science, forthcoming.

6. Chen, K.-Y. and C. R. Plott (2002). Information aggregation mechanisms: Concept, design and implementation for a sales forecasting problem. California Institute of Technology Social Science Working Paper 1131.

7. Cowgill, B. and E. Zitzewitz (2013). Corporate prediction markets: Evidence from Google, Ford and Firm X. Working paper.

8. Dohmen, T., A. Falk, D. Huffman and U. Sunde (2010). Are risk aversion and impatience related to cognitive ability? The American Economic Review 100(3): 1238-1260.

9. Dohmen, T., A. Falk, D. Huffman, U. Sunde, J. Schupp and G. Wagner (2011). Individual risk attitudes: Measurement, determinants and behavioral consequences. Journal of the European Economic Association 9(3): 522-550.

10. Fischbacher U (2007). zTree: Zurich toolbox for ready-made economic experiments. Experimental Economics 10(2): 171-178.

11. Forsythe, R., F. Nelson, G. R. Neumann and R. Wright (1992). Anatomy of an experimental political stock market. American Economic Review 82(5): 1142-1161.

12. Forsythe, R., T. R. Palfrey and C. Plott (1982). Asset valuation in an experimental market. Econometrica 50(3): 537-567.

13. Gillen, B. J., C. R. Plott and M. Shum (2013). Inside Intel: Sales forecasting using an information aggregation mechanism. Working paper.

14. Goel, S., D. M. Reeves, D. J. Watts and D. M. Pennock (2010). Prediction without markets. Proceedings of the ACM EC'10 Conference, Cambridge, MA.

15. Goldstein, D. G., R. P. McAfee and S. Suri (2014). The wisdom of smaller smarter crowds. Proceedings of the ACM EC'14 Conference, Stanford, CA.

16. Greiner, B. (2004). An online recruitment system for economic experiments. In: K Kremer and V Macho (eds.): Forschung und wissenschaftliches Rechnen 2003. GWDG Bericht 63, Göttingen: Ges. für Wiss. Datenverarbeitung, pp. 79-93.

17. Hanson, R., R. Oprea and D. Porter (2006). Information aggregation and manipulation in an experimental market. Journal of Economic Behavior and Organization. 60: 449-459.

18. Healy, P. J., S. Linardi, J. R. Lowery and J. O. Ledyard (2010). Prediction markets: Alternative mechanisms for complex environments with few traders. Management Science 56(11): 1977-1996.

19. Hellinger, E. (1909). Neue Begründung der Theorie quadratischer Formen von unendlichvielen Veränderlichen. Journal für die reine und angewandte Mathematik 136: 210-271.

20. Hopman, J. W. (2007). Using forecasting markets to manage demand risk. Intel Technology Journal 11(2): 127-135.

21. Lichtendahl, K. C., Y. Grushka-Cokayne and R. L. Winkler (2013). Is it better to average probabilities or quantiles? Management Science 59(7): 1594-1611.

22. Offerman, T., J. Sonnemans, G. van de Kuilen and P. Wakker (2009). A truth-serum for non-Bayesians: Correcting proper scoring rules. The Review of Economic Studies 76(4): 1461-1489.

23. Plott, C. and S. Sunder (1982). Efficiency of experimental security markets with insider information: An application of rational-expectations models. Journal of Political Economy 90(4): 663-698.

24. Plott, C. and S. Sunder (1988). Rational expectations and the aggregation of diverse information in laboratory security markets. Econometrica 56(5): 1085-1118.

25. Prelec, D. (2004). A Bayesian truth serum for subjective data. Science 306(5695): 462-466.

26. Schlag, K. and J. van der Weele (2009). Efficient interval scoring rules. Working paper.

27. Smith, V. (1962). An experimental study of competitive market behavior. Journal of Political Economy 70(2): 111-137.

28. Veiga, H. and M. Vorsatz (2006). Price manipulation in an experimental asset market. European Economic Review 53: 327-342.

29. Winkler, R. (1971). Probabilistic prediction: Some experimental results. Journal of the American Statistical Association 66(336): 675-685.

A Experimental instructions

Welcome

You are about to participate in a session on individual decision-making. Thank you for agreeing to take part. The session should last 60 to 90 minutes.

You should already have turned off all your mobile phones, smart phones, mp3 players and any such devices. If not, please do so immediately. These devices must remain switched off throughout the session. Place them in your bag or on the floor besides you. Do not have them in your pocket or on the table in front of you.

The entire session will take place through the computer. You are not allowed to talk or to communicate with other participants in any other way during the session.

You are asked to abide by these rules throughout the session. Should you fail to do so, we will have to exclude you from this (and future) session(s) and you will not receive any compensation for this session.

We will start with a brief instruction period. Please read these instructions carefully. They are identical for all participants in this session with whom you will interact. If you have any questions about these instructions or at any other time during the experiment, then please raise your hand. One of the experimenters will come to answer your question.

Compensation for participation in this session

In addition to the 3.00 Euro participation fee, what you will earn from this session will depend on your decisions and chance. In the instructions and all decision tasks that follow, payoffs are reported in Experimental Currency Units (ECUs). At the end of the experiment, the total amount you have earned will be converted into Euros using the following conversion rate:

1 ECU = 6 Eurocents.

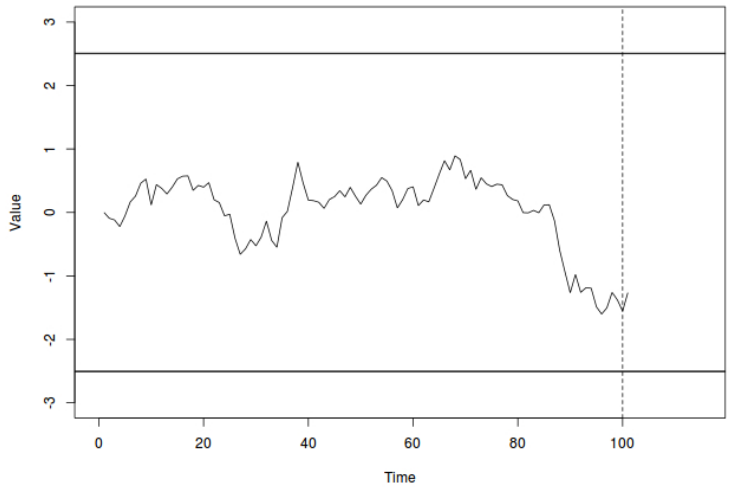
The payment takes place in cash at the end of the experiment. Your decisions in the experiment will remain anonymous.

Instructions

This session consists of twenty rounds. Each round you are faced with two decision tasks and the payoff (in ECU) that you collect depends on the decisions you make and chance. At the end of the session you are paid according to eight random draws (with replacement) over the

payoffs you earned over the two tasks in the twenty rounds.[5]

Before the first round starts, you will be shown the *time series* that results from some *random process*. See the figure below for an example of such a time series.



The random process from which the time series has been generated is kept fixed during the entire session, but every round a different time series will be generated using the same random process. Each round you will see a new time series; so, you will get better acquainted with the random process over rounds. Apart from the realized time series in the previous rounds and the time series shown to you at the beginning (and the one in the figure above), no further information will be given, except that the time series will start at a value of 0 at time $t = 0$.

Each round, before you see the time series that is generated for that round, you are faced with two prediction tasks:

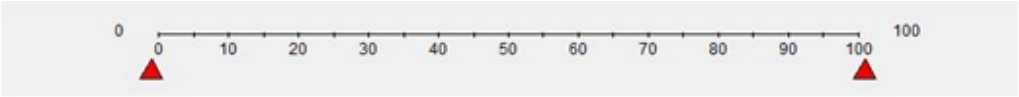
1. First, you are asked how likely you regard the event that the time series hits the *boundary* (one of the thick horizontal lines in the figure above) before time $t = 100$. You can express your expectation regarding this event by moving the triangular cursor along the line. See the figure below.

   The payoff that you earn with this decision task depends on the point you select along the line and the generated time series. The potential payoffs in the event that the time

[5]To elaborate, in total you make 40 decisions that lead to 40 payoffs. From these 40 payoffs, eight are drawn for actual payment. These draws are taken with replacement, meaning that it is not excluded that the same payoff is drawn multiple times, and for each participant individually.

series hits the boundary before time $t = 100$ and in the event that it does not are shown on-screen in real-time when you move the cursor along the line.

2. Second, conditionally on the time series hitting the boundary before time $t = 100$, you are asked to indicate within which *time interval* you think the time series will hit the boundary. You can indicate this interval by moving two triangular cursors (one indicating the lower bound of the interval; the other indicating the upper bound of the interval) along the time line. See the figure below.



   Only in the event that the time series hits the boundary within the indicated time interval you collect a payoff. The smaller the interval that you indicated, the larger this potential payoff is. This potential payoff is shown on-screen in real-time when you move the cursors along the time line.

3. To avoid the unfortunate event that you confirm your decisions while not being completely confident these being the right decisions, you have to approve your decisions at the bottom of the screen.

After having made your predictions, the time series generated for that round will be shown to you. Furthermore, you are informed about the payoffs you collected. It is important to note here that the time series is generated by a statistical software package and is not manipulated for the purpose of this experiment. As all time series shown to you are generated from the same random process, over rounds you will gradually become more familiar with the underlying process.